Phase vocoder and beyond

Marco Liuni and Axel Röbel

(Received January 4, 2013)

The term *vocoder*, which refers to the coding of a voice's features in order to reproduce it synthetically, was introduced in a work by Dudley [Dud39]: there, a first system of speech analysis and re-synthesis was presented, together with the first technical implementation of sound models that have been later improved and largely exploited (see section 2.1).

Then, the *phase vocoder* was originally introduced in 1966 by Flanagan [FG66], working at the Bell Labs. This technique is based on the STFT (Short Time Fourier Transform, see section 1): it began to be widely exploited when the dedicated algorithms were made computationally fast enough (see [CT65] for the Fast Fourier Transform original algorithm, and [Por76] for an implementation of STFT taking advantage of the FFT). The input of the STFT is a generic sound, the output is a set of coefficients that allow a perfect reconstruction of the original sound in terms of atomic signals, which are weighted modulated sinusoids. A main advantage, in relation to previous techniques such as the *Heterodyne filter* (see the dedicated chapter in [Moo75]), is that no knowledge about the fundamental frequency is needed, thus making the method well suited for a broad range of sounds. On the other hand, the representation is not related to a sinusoidal decomposition, so that the sound's sinusoidal components have to be deduced from the coefficients by means of *sinusoidal modeling* techniques (see section 2.1). Despite of some drawbacks, namely a given unavoidable small latency, as well as artifacts that in certain cases are introduced in the transformed sound, an increasing range of high-quality sound processing techniques are currently based on the phase vocoder and its improvements (see [LD99a; LD99b] and the related bibliographies): among the most popular, the variation of a sound's duration without affecting its pitch (also known as *time stretch*) and the shift of a sound's pitch without changing its duration.

1 Basic concepts about the phase vocoder

Time-frequency representations (often indicated as time-frequency distributions, see [Coh95; Coh89; Mal99] for the theory and the motivations beyond this approach), briefly indicated as *TFR*, are employed for several different signals: sound, light, image, video, and other phenomena that are interpretable as a function with finite energy on a real or complex space. The starting point of this prolific field is the work of the french mathematician and physicist Jean Baptiste Fourier, together with the improvements of computer science techniques for the fast application of models and tools stemming from his results. The first goal of a signal representation is to increase its readability: the spectrum of a sound is a fundamental characterization of its features in the frequency domain, but it is not enough to have a complete local information. If we consider a signal and its Fourier transform separately, we cannot observe the evolution of its spectral content over the time. With TFRs, a further characterization is provided, increasing the dimension of the representation domain: for a mono-dimensional signal, a TFR is a two-dimensional set that jointly describes its time and frequency content. In the case of STFT, which is the TFR directly related to the phase vocoder, the time localization is obtained by means of the so-called *window* function, which selects a small slice of signal before computing the FFT.

Whereas most of the concepts in this work are treated without mathematical rigor, here some fundamental mathematics are recalled, adopting the notation used by Laroche in [LD99a]: they are needed to precisely describe the phase vocoder scheme, composed of the three steps analysis/transformation/re-synthesis. Let x be a real-valued discrete signal, $x \in \mathbb{R}[\mathbb{Z}]$, and h a symmetric real-valued discrete signal composed of N samples, $h \in \mathbb{R}^N$. In the analysis stage, the time step (the *hop size*) is $R_a \in \mathbb{N}$ while the time index is indicated as t_a; having $u \in \mathbb{Z}$, the u-th time position corresponds to $t_a^u = R_a u$. The discrete STFT of x with window h is a discrete FFT of x multiplied by a time-shift of h, given by

$$X(t_a^u, \Omega_k) = \sum_{n \in \mathbb{Z}} h(n)\, x(t_a^u + n)\, \mathrm{e}^{-j\Omega_k n}\,, \tag{1}$$

where $k = 0, ..., N-1$ and $\Omega_k = \frac{2\pi k}{N}$ is the frequency variable. The above definition shows that X is a two-dimensional complex-valued representation; as such, its coefficients can be expressed in terms of real and imaginary part, or amplitude and phase: this second form provides an interpretation of the coefficients that is well-suited for sinusoidal signal models, and is therefore adopted. The amplitude of the coefficient at time position a and frequency bin k is given by $|X(t_a^u, \Omega_k)|$, while its phase is $\angle X(t_a^u, \Omega_k) = \arg(X(t_a^u, \Omega_k))$.

In order to analyze the characteristics of a sound, and to modify them introducing different desired qualities, the coefficients of a time-frequency representation need to be interpreted within an appropriate sound model (see section 2.1). A transformation

$\mathrm{T}(X) = Y$ may then be performed in the time-frequency domain, leading to a complex-valued representation Y of the same dimension as X. Finally, a signal y can be synthesized from the transformed TFR using the inverse procedure of the one described above: as done for analysis, let $R_s \in \mathbb{N}$ be the synthesis hop size, t_s the time index and $t_s^u = R_s u$ the u-th time position. Then, first setting

$$y_u(n) = \frac{1}{N} \sum_{k=1}^{N-1} Y(t_s^u, \Omega_k)\, \mathrm{e}^{j\Omega_k n} \,, \tag{2}$$

and using the same window function h, the synthesized signal is given by

$$y(n) = \sum_{u \in \mathbb{Z}} h(n - t_s^u)\; y_u(n - t_s^u) \,. \tag{3}$$

2 Improvements and extensions

The possibility to shift between the signal and its STFT, that is between the time domain and the time-frequency domain, is a first advantage of the framework described above: replacing Y by X in (2) with $t_s^u = t_a^u$, equation (3) provides a perfect reconstruction of the original signal x using its STFT, under certain constraints for the analysis parameters [GL84]. Nevertheless, a given representation Y is not necessarily the STFT of any sequence $y \in \mathbb{R}[\mathbb{Z}]$ (see [Mal99], Proposition 4.1 for a necessary and sufficient condition in the case of continuous signals): this means that, given a complex-valued representation Y of the appropriate dimension, by computing y according to equation (3), the STFT of y does not necessarily coincide with Y.

A direct consequence of the previous remark is that a modified STFT $Y = \mathrm{T}(X)$ of an original sound x may not coincide with the STFT of any existing sound; in this sense, equation (3) provides a signal y whose STFT is an approximation of Y. An alternative reconstruction formula is proposed in [GL84], which guarantees that the STFT of the reconstructed signal y gives the best approximation of the initial representation Y, in the sense of the mean squared error. This formula is now commonly adopted within high-quality phase vocoders. Moreover, it shows important analogies with the reconstruction formulas known as *painless non-orthogonal expansions* (see [DGM86] for the original formulation), in the context of Gabor frames theory (see [Grö01] for a complete survey, and section 3.1 for some links with the phase vocoder).

2.1 Sound models

As pointed in [Röb10a], an efficient signal model should use perceptually relevant components, that have a simple relation with the physical properties of the sound sources. The simpler the relation between the perceptually relevant properties

of the physical sound source and the signal model, the easier it should be to provide controls that reflect our intuition, that is built on physical interaction.

This kind of relations exist for example for models that are represented in terms of the vibration modes; these individual modes can be represented in a rather simple manner, as a sinusoid with time-varying amplitude and frequency. In the case of analysis/re-synthesis systems the modal representation is achieved by means of the sinusoidal model [SS90; Ser97; RS07]. The vibrating modes, however, are generally not sufficient to describe a given sound signal. Noise sources are present in nearly all cases, for example as a side effect of the excitation. Generally one assumes that the noise is independent from the sinusoidal components and a noise component is added into model without destroying the simplicity of the transformation.

A mathematical formulation of the sinusoids plus noise model that has been discussed is

$$\begin{array}{rcl} p_k(n) & = & a_k(n)\cos(\Theta_k(n)) \\ s(n) & = & \sum_k p_k(n) + r(n). \end{array} \tag{4}$$

Here p_k is a sinusoid with time varying amplitude $a_k(n)$ and phase $\Theta_k(n)$ and $s(n)$ is the signal that consists of a superposition of sinusoids and a noise component $r(n)$. Because all components of the sinusoidal model can independently vary amplitude and frequency over time, the model allows representing onsets, vibrato and other signal modulations.

2.1.1 Sinusoids plus noise model

The sinusoidal models have their origin in the vocoder developed by Dudley in 1939 [Dud39]: his ideas evolved with the invention of computers and digital signal processing into early versions of the phase vocoder [FG66]. These phase vocoders used very low number of bands (30 bands with 100Hz bandwidth) such that the resolution of the individual sinusoids could not be guaranteed. With further increasing computing capacities and the use of FFT algorithms the number of bands (today bins) increased and as a next step explicit harmonic sinusoidal models were developed [MQ86]. The use of the sinusoidal modeling techniques for musical applications also started with the early phase vocoder [Moo78] and evolved into an explicit sinusoidal model [SS87]. The main advantage of the explicit sinusoidal model compared to the phase vocoder was the peak picking that was part of the analysis for the explicit sinusoidal models. The peak picking and subsequent parameter estimation did allow to increase frequency resolution and improved the tracking of time varying sinusoids. As a next step the sinusoidal model was extended by means of a dedicated noise model [SS90] so that the sinusoidal model in (4) was completed. After the introduction of the intra-sinusoidal phase synchronization ([LD99a], see section 2.2) the phase vocoder has evolved into an implicit implementation of a sinusoidal model that generally is computationally more efficient than the explicit sinusoidal model. Due to the fact that the phase vocoder representation achieves

a better representation of potential structure in the aperiodic (noise) component, it often achieves better quality than the explicit sinusoidal model.

The main problem with the sinusoids plus noise model is related to finding the model parameters from the original signal. This problem has triggered numerous research efforts over the last decades. Despite the many interesting and powerful methods that have been developed, and that extended the boundaries of the signal representation that can be obtained using this model, there are still open problems (noise models are discussed in section 3.2).

2.1.2 Source-filter model

The source-filter model is another important signal model that is widely used for signal transformation algorithms. It has the same origins as the sinusoids plus noise model [Dud39]. In its first application, the excitation source had been represented by either an impulse train parametrized by the fundamental frequency, or by means of white noise. In both cases, the filter part has been achieved by means of modulating the energy of the excitation signal in bands of constant bandwidth (=250Hz). This basic setting is still in use today. The band wise filtering will in most cases be replaced by a continuous filter function that is called the spectral envelope [MG82; RS07].

The source filter model has many applications for signal transformations. Cross synthesis for example can be achieved by means of using the excitation signal (source) from one signal and the resonator (filter) from another. Other applications are transformations that require independent transposition of pitch and formant structure.

An important precondition for the source filter model is that the distinct source and filter parts can be estimated from the original signal. One of the first techniques that has been used for spectral envelope estimation is linear prediction (LPC) [Mak75]. This method assumes an autoregressive filter function. It has been used especially for speech signals, for which the autoregressive filter model has a physical justification, at least for some configurations of the vocal tract [MG82]. A problem of the LPC estimate is the fact that it is strongly biased if the excitation spectrum contains sinusoids. This problem has been addressed in the discrete all pole model [EJM91]. Alternative spectral envelope estimators use the cepstral representation to derive the spectral envelope. An early rather costly and complex method is the discrete cepstrum [CM96]. Later a more efficient method has been developed [RR05; RVR07] that is using the same envelope representation but makes use of a rediscovered proposal of an iterative cepstral envelope estimator [IA79]. The method is referred to as *True Envelope Estimator*. It has proven to provide nearly optimal estimates given that the spectral envelope can only be observed in a strongly sub sampled version that is produced by the sinusoidal components sampling the filter transfer function [VRR06].

The estimation of the noise envelopes of the background noise that is part of a complex sound consisting of sinusoidal and noise components is a problem that has received relatively few interest. The estimation of the sinusoidal parameters and estimation of the noise level from the residual is a possible procedure, but if the sinusoidal components are superimposed as for example in polyphonic music this procedure will not provide robust results. There exist only a few methods that allow to establish a background noise estimate for complex polyphonic sounds [MHM06; YR06].

The source-filter model can take advantage of a sound decomposition in sinusoidal and noise components: two different filters can be estimated for the two parts, allowing a refined controls of the appropriate parameters. In the context of instruments modeling, for instance, such an extension has been developed in models that represent an instrument by means of time-varying source and filter [Kla07]; or by means of source and filter depending on the pitch and the intensity of the played note, ideal for extended samplers [HR12].

2.2 Phase coherence and transient preservation

The standard phase vocoder performs signal transformation by means of modifying and moving the spectral frames of an STFT analysis of the sound to be transformed [Ser97; LD99a]. During transformation, the spectral frames $y_u(n)$ in equation (2) are modified in content and position [LD99a; Röb03], yielding a sequence $\widetilde{y_u}(n)$ that is then synthesized using the reconstruction formula in (3), or other overlap-add techniques as the one in [GL84], discussed in section 2. Whenever the STFT frames are time-shifted, which means that the synthesis frame position t_s^u is different from t_a^u, the phases of the STFT have to be adapted to achieve coherent overlap-add of the sinusoidal components. Within the phase vocoder this phase adaptation is based on the observed phase evolution in all the bins of the original signal frames. Phases at position t_s^u are obtained from phases at position t_s^{u-1}: being based on the evolution of the phase of individual bins along time, this feature is referred as *horizontal phase synchronization*.

The phase synchronization along time alone does not guarantee that the features of the sound components are correctly preserved in a transformation: for instance, the different bins involved in the representation of a single stationary sinusoid could be not synchronized after a time-stretch, causing the transformed sound to be different from a sinusoid. The extension of this basic example to the different components of a sound makes clear that its timbre may not be necessarily preserved with such a transformation. In [Puc95], a first trial addressing this problem was made. A major improvement in this sense has been the introduction of a method to preserve *vertical* phase synchronization [LD99a].

Even with the phase update just discussed, the phase vocoder does not take into account the phase relations between the different sinusoids. Therefore, frequency estimation errors will result in a desynchronization of the different sinusoidal

components. While the vertical de-synchronization of the sinusoidal components is perceptually uncritical for most musical signals, for speech signals it affects the perception of the underlying excitation pulses, and leads to an artifact that is generally described as missing clarity (phasiness) of the transformed voice. To overcome this problem, different strategies for *shape invariant processing* have been introduced [QM92; Röb10b], denoting transformation algorithms that preserve these inter-partial phase relations.

If the amplitude of a sinusoid changes abruptly, a situation that arises for example during attack transients or note onsets, the prerequisites of the phase correction are no longer valid. Consequently, the results obtained with the phase vocoder have poor quality. Time stretching attack transients, with the phase vocoder, results in less severe cases in softening of the perceived attack. In more severe cases a complete change of the sound characteristics may take place.

Influenced in part by the sinusoids with noise and transients model proposed in [LSI98], some solutions have been tried, based on detecting transient time positions, and reinitializing the phase at these transient positions [Bon00; DDS02]. The argumentation is straightforward: the phase update algorithm of the phase vocoder tries to ensure phase coherence between the current and previous frame, which in case of an onset or other transient events is not appropriate. In the transient aware phase vocoder algorithms, this had been accomplished by means of detecting transients, reinitializing the phase for the detected regions and forcing the time stretching factor to be one during the transient regions. The transient detection is usually based on energy change criteria in rather broad bands and the phase is reinitialized for all bins in the frequency band detected as transient.

Unfortunately there exist fundamental problems with these approaches. Reinitialization of all phases in a transient segment would certainly destroy the phase coherence of stationary partials from other sound sources that might exist in the segment tht is considered to be transient. Moreover, fixing the delay factor to one in the transient regions requires automatic compensation in non transient regions to achieve the overall requested stretch factor. For a dense sequence of transients this may be difficult to achieve. Accordingly, an improvement of the mentioned approaches is proposed in [Röb03], such that an existing transient or onset event could be handled in a more local manner. The goal of the refined determination of transients is to be able to reset the phase only for these spectral peaks and to achieve the transient reinitialization, without requiring a local change of the time stretching factor.

3 Perspectives and ongoing research

The quality of techniques based on the phase vocoder is deeply linked to the representation and sound model adopted. In this section, two main research topics

are detailed, whose results may contribute to the improvement of the existing sound processing methods that are based on the phase vocoder: some extensions of the STFT are considered, which extend the class of TFRs that the phase vocoder refers to, as well as an insight into the noise part of the sinusoidal sound model.

3.1 Variable resolution and adaptivity

STFT closely reflects the concept of time-varying spectrum: its coefficients allow a direct interpretation in terms of amplitude and frequency of the sound components they refer to. Nevertheless, the need to define a window function introduces a dependance of the transform on the window used; in particular, since the window is fixed, the STFT has constant resolution over the whole time-frequency plane. This is a limit, as the precision needed to separate the information coming from different components of complex sounds may vary significantly.

As a basic example, a percussion sample with fast sequences of transients can be considered, that one may want to fit for a different tempo than the original one. If the support of the analysis window is too large, it is possible that a given time-shift includes several transients. From the analysis point of view, these components are indivisible, which means that every treatment concerning their analysis frame applies to them all: in the case of a time-stretch of the original sample, this is particularly inappropriate, because it makes impossible to situate different transients independently.

A symmetric basic example is the case where a small analysis window is used with instruments having close partials: as the frequency resolution of a window function is directly proportional to the size of its support, in this case the value of a frequency bin in the analysis may be influenced by different partials. This degrades the accuracy of spectral processing techniques, like a pitch-shift.

The problem of conceiving TFRs with variable time-frequency resolution is fundamental for many different domains: analyses with a non-optimal resolution lead to a blurring, or sometimes even a loss of information about the original signal, which affects every kind of later treatment. This motivates the research for adaptive methods, currently conducted in both the signal and the applied mathematics communities: they lead to the possibility of analyses whose resolution locally changes according to the signal features. Here, some approaches providing a direct application to the phase vocoder are considered.

Going back to the previous examples, the size of the window function is shown to be peculiar for two main reasons: it determines the time precision of the transform, as well as its frequency resolution, the two being inversely proportional; therefore, an STFT highly localized in time could not have a high frequency resolution, and vice versa. A possible strategy to partially overcome this drawback is to consider analyses with variable window size: a better time or frequency resolution can be privileged at different time positions within a same sound analysis. This adaptation

can be made automatically, by the choice of a specific window within a given set, according to appropriate rules, or measures [RBW10; Liu+10; Bal+11].

The referenced approaches can be interpreted in the context of Gabor frames theory (see [Grö01] for a complete survey): for all of them, there exist reconstruction formulas extending the one in equation (3), that provide perfect reconstruction of the original signal. Moreover, if certain conditions about the density of the analysis discretization and the windows used are fulfilled, these formulas reduce in a highly efficient form: that is, the principal computational cost of the related algorithms are due to the number of FFTs that are performed.

IRCAM's SuperVP[1] library is the only software, to the authors' knowledge, allowing the variable-window extension to the standard phase vocoder. A recent perceptive test (to appear in [Liu+13]), conducted by the authors, has compared the quality of standard time-stretches of several sound files, to that obtained with the automatic adaptation of the window size proposed in [Liu+10]: for most of the sound files, this test has shown a significant increase of the quality given by a variable adaptive window, in relation to the one achieved by choosing a different fixed window for each sound file.

The variable-window approach extends the STFT in the direction of a non-uniform sampling in time: that is, the time index t_a^u in equation (1) (and t_s^u, accordingly) does not vary linearly, but is linked to the window size used at each specific time location. A symmetric approach, still providing perfect reconstruction of the original signal with efficient algorithms, consists in extensions of the STFT that allow non-uniform sampling of the frequency dimension: that is, the STFT bins' frequencies, Ω_k in equation (1), are not equally spaced, and the modulation of the analysis window as well as the hop size are frequency-dependent. This approach allows the design of phase vocoders with arbitrary frequency band selection (see [EDM12], which generalizes previous works mentioned within the references).

A further step consists in the local adaptation of the STFT resolution depending on both the time and frequency positions of the coefficients [JT07; Dör11; LBR11]: with such methods, time or frequency precision of the analysis can be independently specified, according to the local features of the input sound. Currently, the main open problems concern two aspects: first, for this case, algorithms providing perfect reconstruction are in general not efficient. Therefore, to guarantee a low computational cost of the global analysis/re-synthesis stage, an approximation error has to be considered. Moreover, while traditional processing techniques can be extended to the case of time-adapted window with limited efforts, if a single analysis frame is obtained by the modulations of different window functions and variable spacing of the frequency bins, then a global re-definition of the sinusoidal modeling and the processing within such an extended phase vocoder is needed.

[1] SuperVP is a library providing an extended phase vocoder; see <http://anasynth.ircam.fr/home/english/software/SuperVP>.

3.2 Preserving statistical properties of a sound texture

Being based on sinusoidal models, the phase vocoder does not provide satisfying results when applied to non-sinusoidal signals, or sound textures like passing cars, wind blowing, or applauses: these are sounds that are hard to represent with a sinusoidal model, and are better described by means of statistical descriptors of the signal spectrum. The perceptual relevance of a number of statistical descriptors for the recognition of sound textures has recently been shown in [MS11]. These results are highly interesting for the present topic because they indicate that similar to the parameters amplitude and frequency, that characterise the perceptually relevant properties of a sinusoidal component, for noises there exist perceptually relevant parameters as well. Similar as for sinusoidal components, these statistical descriptors have to be preserved during time stretching for the texture to remain the same.

These ideas have triggered initial investigations to test the use of these statistical descriptors in STFT-based signal transformation algorithms. In [LRWS12], some relevant statistical properties of the time-frequency representation of white noise have been investigated: an algorithm is proposed, that performs time-stretch of noise preserving essentially the autocorrelation of the complex signal in the STFT bins: this is shown to be one of the quantities, that are needed for the result to be perceived as a noise of the same type. There are only rather few statistical parameters that are relevant for the perception of white noise, and those can be preserved relatively easily. But for the more general case of time stretching sound textures, the appropriate selection of statistical descriptors, as well as the algorithms that should be used to preserve them, are subject of ongoing research.

4 The phase vocoder in Marco Stroppa's *Zwielicht* (1994-99)

Time-frequency analysis is a natural context for the modeling of time-evolving spectra, thus in particular for sounds and music. One of the interest of time-frequency analysis and processing is to establish relations between sounds themselves: a sound representation makes it easier to define and work with classes of timbre, and to visualize sound components. Once defined a target sound or effect to realize, such knowledge is a useful tool for the orientation among the large range of processing and re-synthesis methods available. Several composers have developed a deep musical experience of these techniques. Among them, the approach of Marco Stroppa (this section is based on an interview with the composer, some of his sentences are quoted) gives a special outlook on the musical potential offered by a complete framework for sound analysis, manipulation and re-synthesis: "Ho appreso *SuperVP* come se fosse uno strumento; ho imparato ad usarlo, e questo mi ha dato la coscienza di cosa poterci fare"[2].

[2] "I've learned SuperVP like an acoustic instrument; I've learned how to use it, and this gave me the knowledge about what I can do with it."

In his work *Zwielicht* (1994-99, for double bass, two percussions, electronics and 13-D sound projection), most of the electronic material is obtained with *AudioSculpt*[3], in the 1.2 beta 1 version available in 1996, and with the command line SuperVP when the parameters of the transformation had to vary dynamically.

Zwielicht means "between two lights", and the title refers to sounds staying at the frontier of the cognitive world of an instrument, without leaving it. The work lasts 35 minutes, and is composed of 13 tracks for a 13-loudspeakers sound projection, one being hidden above the ceiling. The electronic sounds, which are uniquely obtained by processing samples of metallic percussions and double bass, are divided in families that represent cognitive units. Such units refer to the capability of recognizing that a given sound profile belongs to the same family, even when its generating event is altered: "Il processo cognitivo si fonda sulle emergent properties, che risultano dai rapporti tra parametri, oltre che dai valori dei parametri in sé"[4].

The sounds that are used are at the frontier of cognition in two senses: on one side there are those with minimal intensity, so weak sounds that they cannot be heard a few centimeters away from their source, like a knitting needle gently scrubbing the edge of a crotale. For those sounds, the recording method is itself a compositional choice, as different methods can lead to completely distinct results. And still, no matter which method is adopted, an extreme pre-amplification is needed for the sound to become audible, introducing a strong noise that has to be processed further.

With a different meaning of frontier, there are events whose production is made through an unstable process, that are therefore not systematically reproducible, without being necessarily weak: for example, the harmonic of a cymbal held with the fingers, and scrubbed with a bow. Such a sound is unstable, and recording is a way to set it, making it stable.

Some of the treatments conceived by the composer are detailed here, those in particular where the potential of an advanced phase vocoder is made clear. A first example is transposition, with or without altering the original duration: it is applied to the unstable sounds, where the pitch is constant once the fundamental gets stable. They are transposed in order to have the same profile at desired pitches, as the sound event itself cannot be mechanically tuned. The composer reports that transposing at intervals larger than a fourth-fifth did not provide a satisfying quality, with the SuperVP version available at that time: the advancing concerning the vertical phase synchronization of the STFT (detailed in section 2.2) as well as the transformation with spectral envelope preservation (see section 2.1.2) have been introduced later, and currently allow transpositions that sound natural even at larger intervals.

Another example shows how the composer adapted his musical ideas to a special feature of SuperVP, in order to find their most suitable realization. This feature implements a spectral binary mask, that is a particular filter bank made of alternating

[3] AudioSculpt is a software for viewing, analysis and processing of sounds, based on the SuperVP library; see <http://anasynth.ircam.fr/home/english/software/audiosculpt>.

[4] "The cognitive process is based on the emergent properties, that issue from the relations between parameters, and also on the parameters' values themselves".

band-pass and band-stop filters with an almost vertical slope. After specifying whether the first band is stopped or passed, a simple list of frequencies defines the borders of successive bands. In one case, the composer chose to compute narrow pass-bands centered around the spectral components of a given fundamental frequency. When this filter is applied to sounds with the same fundamental frequency, it highlights their spectral characteristics, thus producing a sort of noise reduction. However, when it is applied to sounds with a different fundamental frequency, the filter bank tends to fall on low-energy parts of the spectrum, outside of the main peaks. In this case, the floating-point precision of SuperVP allows to normalize the result, in order to make it audible without loosing sound quality. One has a sort of "hollow" shadow of the original sound. This process was also used dynamically: the frequencies of a filter bank are interpolated over time, according to data specified in a text file written by the composer and passed to SuperVP as an argument. The only limitation is that the overall amount of bands and the nature of the first band cannot change during one process. When applied to a sound with a single pitch, this allows to explore its spectral contents and to cross regions of different amplitude, thus generating a lively spectral glissando effect with amplitude swells.

A last treatment, which is detailed here, is based on the principle of altering the internal rhythm of a sound, imposed on it through the recording technique: for instance, a triangle hung up with wrapped support, which is hit, then released and sampled while it unwraps, obtaining a slowed down flanger effect. This sound has been reversed, to realize an accelerando and a crescendo, then transposed 6 times, changing the duration too, so that the transposition interval and the rhythm are correlated. The result has finally been synchronized with a *temporal pivot* [DS90], in order to temporally align the climax of the crescendo.

Stroppa reports that, when using treatments based on the phase vocoder, "apprendistato e composizione sono due fasi diverse"[5]: the instrument exploration stage, in relation with the sound material, needs ergonomic interfaces, and the possibility to listen quickly to the results of the processing. But those needs may reduce the results' quality, since the parameters for analysis and processing are set in order to give priority to the computational speed. In the following stage of composition, once the treatments have been selected, priority has to be given to the highest expressive power of processing, and precision in following the changes of parameters that vary dynamically. In this sense, his choice for the appropriate software for his work is driven by two criteria: quality first, but also the possibility to control the engine dynamically, according to the score he conceived, and evaluating its results.

[5] "Training and composition are two different stages".

References

[Bal+11] P. Balazs et al. “Theory, implementation and applications of nonstationary Gabor frames”. In: *Jour. of Computational and Applied Mathematics* 236.6 (2011), pp. 1481 –1496. URL: http://www.sciencedirect.com/science/article/pii/S0377042711004900.

[Bon00] J. Bonada. “Automatic technique in frequency domain for near-lossless time-scale modification of audio”. In: *Proc. of the International Computer Music Conference, ICMC*. 2000, pp. 396–399.

[CM96] O. Cappe e E. Moulines. “Regularization techniques for discrete cepstrum estimation”. In: *Signal Processing Letters, IEEE* 3.4 (1996), pp. 100 –102.

[Coh89] L. Cohen. “Time-frequency distributions - A review”. In: *Proc. of the IEEE* 77.7 (lug. 1989), pp. 941–981.

[Coh95] L. Cohen, cur. *Time-Frequency Analysis*. Upper Saddle River, New Jersey, USA: Prentice-Hall, 1995.

[CT65] J. W. Cooley e J. W. Tukey. “An algorithm for machine calculation of complex Fourier series”. In: *Math. Comp.* 19 (1965), pp. 297–301.

[DDS02] Chris Duxbury, Mike Davies e Mark B. Sandler. “Improved Time-Scaling of Musical Audio Using Phase Locking at Transients”. In: *Audio Engineering Society Convention 112*. Apr. 2002. URL: http://www.aes.org/e-lib/browse.cfm?elib=11324.

[DGM86] I. Daubechies, A. Grossmann e Y. Meyer. “Painless nonorthogonal expansions”. In: *Jour. Math. Phys.* 27 (mag. 1986), pp. 1271–1283.

[DS90] J. Duthen e M. Stroppa. “Une représentation de structures temporelles par synchronisation de pivots”. In: *Proc. of Symposium Musique et Assistance Informatique*. Marseille, France, 1990, pp. 471–479.

[Dud39] H. Dudley. “Remaking Speech”. In: *Jour. of the Acoustical Society of America* 11.2 (1939), pp. 169–177. URL: http://link.aip.org/link/?JAS/11/169/1.

[Dör11] M. Dörfler. “Quilted Gabor frames - A new concept for adaptive time-frequency representation”. In: *Advances in Applied Mathematics* 47.4 (2011), pp. 668 –687. URL: http://www.sciencedirect.com/science/article/pii/S0196885811000157.

[EDM12] G. Evangelista, M. Dörfler e E. Matusiak. “Phase Vocoders With Arbitrary Frequency Band Selection”. In: *Proc. of the Int. Conf. Sound and Music Computing, SMC12*. 2012, pp. 153–178.

[EJM91] A. El-Jaroudi e J. Makhoul. “Discrete all-pole modeling”. In: *IEEE Trans. Signal Processing* 39.2 (1991), pp. 411 –423.

[FG66] J. L. Flanagan e R. M Golden. “Phase vocoder”. In: *Bell System Technical Journal* 45 (1966), pp. 1493–1509. URL: http://ci.nii.ac.jp/naid/30016140882/en/.

[GL84] D.W. Griffin e J.S. Lim. “Signal Estimation from Modified Short-Time Fourier Transform”. In: *IEEE Trans. Acoust. Speech Signal Process.* 32.2 (apr. 1984), pp. 236–242.

[Grö01] K. Gröchenig, cur. *Foundations of Time-Frequency Analysis*. Boston, Massachusetts, USA: Birkhäuser, 2001.

[HR12] Henrik Hahn e Axel Röbel. “Extended Source-Filter Model of Quasi-Harmonic Instruments for Sound Synthesis, Transformation and Interpolation”. In: *Proc. of the Int. Conf. on Sound and Music Computing (SMC 2012)*. Copenhagen, Denmark, 2012. URL: http://articles.ircam.fr/textes/Hahn12a/.

[IA79] S. Imai e Y. Abe. “Spectral envelope extraction by improved cepstral method”. In: *Journal of IEICE* 62 (1979), pp. 217–223.

[JT07] F. Jaillet e B. Torrésani. “Time-frequency jigsaw puzzle: adaptive and multilayered Gabor expansions”. In: *Int. Jour. for Wavelets and Multiresolution Information Processing* 1.5 (2007), pp. 1–23.

[Kla07] A. Klapuri. “Analysis of Musical Instrument Sounds by Source-Filter-Decay Model”. In: *Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP07*. Vol. 1. 2007, pp. I–53 –I–56.

[LBR11] M. Liuni, P. Balazs e A. Röbel. “Sound analysis and synthesis adaptive in time and two frequency bands”. In: *Proc. of DAFx11*. Paris, France, 2011.

[LD99a] J. Laroche e M. Dolson. “Improved phase vocoder time-scale modification of audio”. In: *IEEE Trans. Speech and Audio Processing* 7.3 (1999), pp. 323 –332.

[LD99b] J. Laroche e M. Dolson. “New phase-vocoder techniques for pitch-shifting, harmonizing and other exotic effects”. In: *Applications of Signal Processing to Audio and Acoustics, 1999 IEEE Workshop on*. 1999, pp. 91 –94.

[Liu+10] M. Liuni et al. “A reduced multiple Gabor frame for local time adaptation of the spectrogram”. In: *Proc. of DAFx10*. Graz, Austria, 2010, pp. 338 –343.

[Liu+13] M. Liuni et al. “Automatic Adaptation of the Time-Frequency Resolution for Sound Analysis and Re-Synthesis”. Accepted for publication in IEEE Trans. Audio, Speech and Language Processing. 2013.

[LRWS12] W-H. Liao, A. Röbel e A. W.Y. Su. “On stretching Gaussian noises with the phase vocoder”. In: *Proc. of the 15th Int. Conf. on Digital Audio Effects (DAFx-12)*. 2012.

[LSI98] Scott N. Levine e Julius O. Smith III. “A Sines+Transients+Noise Audio Representation for Data Compression and Time/Pitch Scale Modifications”. In: *Audio Engineering Society Convention 105*. Set. 1998. URL: http://www.aes.org/e-lib/browse.cfm?elib=8399.

[Mak75] J. Makhoul. “Linear prediction: A tutorial review”. In: *Proc. of the IEEE* 63.4 (1975), pp. 561 –580.

[Mal99] S. Mallat, cur. *A wavelet tour on signal processing*. San Diego, California, USA: Academic Press, 1999.

[MG82] John E. Markel e A. H. Gray. *Linear Prediction of Speech*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 1982.

[MHM06] Guillaume Meurisse, Pierre Hanna e Sylvain Marchand. “A New Analysis Method for Sinusoids+Noise Spectral Models”. In: *Proc. of the 9th Int. Conf. on Digital Audio Effects, DAFx06*. 2006. URL: http://hal.archives-ouvertes.fr/hal-00307892.

[Moo75] James A. Moorer. “On the Segmentation and Analysis of Continuous Musical Sound by Digital Computer”. Tesi di laurea mag. Stanford, CA: Stanford University, 1975. URL: https://ccrma.stanford.edu/files/papers/stanm3.pdf.

[Moo78] James A. Moorer. “The Use of the Phase Vocoder in Computer Music Applications”. In: *Jour. Audio Eng. Soc* 26.1/2 (1978), pp. 42–45. URL: http://www.aes.org/e-lib/browse.cfm?elib=3293.

[MQ86] R. McAulay e T. Quatieri. “Speech analysis/Synthesis based on a sinusoidal representation”. In: *IEEE Trans. Acoustics, Speech and Signal Processing* 34.4 (1986), pp. 744 –754.

[MS11] Josh H. McDermott e Eero P. Simoncelli. “Sound Texture Perception via Statistics of the Auditory Periphery: Evidence from Sound Synthesis”. In: *Neuron* 71.5 (2011), pp. 926–940.

[Por76] M. Portnoff. “Implementation of the digital phase vocoder using the fast Fourier transform”. In: *IEEE Trans. Acoustics, Speech and Signal Processing* 24.3 (1976), pp. 243 –248.

[Puc95] M. Puckette. “Phase-locked vocoder”. In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, ASSP95*. 1995, pp. 222 –225.

[QM92] T.F. Quatieri e R.J. McAulay. “Shape invariant time-scale and pitch modification of speech”. In: *IEEE Trans. Signal Processing* 40.3 (1992), pp. 497 –510.

[RBW10] D. Rudoy, P. Basu e P.J. Wolfe. “Superposition Frames for Adaptive Time-Frequency Analysis and Fast Reconstruction”. In: *IEEE Trans. Sig. Proc.* Cambridge, Massachussets, 2010, pp. 2581–2596.

[RR05] A. Röbel e X. Rodet. “Efficient Spectral Envelope Estimation and its application to pitch shifting and envelope preservation”. In: *Proc. of the 8th Int. Conf. on Digital Audio Effects, DAFx05*. 2005, pp. 30–35.

[RS07] X. Rodet e D. Schwarz. “Spectral Envelopes and Additive + Residual Analysis/Synthesis”. In: *Analysis, Synthesis, and Perception of Musical Sounds*. A cura di James W. Beauchamp e Robert T. Beyer. Modern Acoustics and Signal Processing. Springer New York, 2007, pp. 175–227. URL: http://dx.doi.org/10.1007/978-0-387-32576-7_5.

[RVR07] A. Röbel, F. Villavicencio e X. Rodet. “On cepstral and all-pole based spectral envelope modeling with unknown model order”. In: *Pattern Recognition Letters* 28.11 (2007), pp. 1343 –1350. URL: http://www.sciencedirect.com/science/article/pii/S016786550700092X.

[Röb03] A. Röbel. “A new approach to transient processing in the phase vocoder”. In: *Proc. of the 6th Int. Conf. on Digital Audio Effects (DAFx03)*. 2003, pp. 344–349.

[Röb10a] A. Röbel. “Between Physics and Perception: Signal Models for High Level Audio Processing”. In: *Digital Audio Effects (DAFx)*. Graz, Austria, 2010. URL: http://articles.ircam.fr/textes/Roebel10a/.

[Röb10b] A. Röbel. “Shape-invariant speech transformation with the phase vocoder”. In: *Proc. of the Int. Conf. on Spoken Language Processing, InterSpeech10*. 2010, pp. 2146–2149.

[Ser97] Xavier Serra. “Musical sound modeling with sinusoids plus noise”. In: *Musical Signal Processing*. A cura di C. Roads et al. Lisse, the Netherlands: Swets & Zeitlinger Publishers, 1997, pp. 91–122.

[SS87] J. O. Smith e X. Serra. “PARSHL: an analysis/synthesis program for non-harmonic sounds based on a sinusoidal representation”. In: *Proc. of the International Computer Music Conference, ICMC*. 1987, pp. 290–297.

[SS90] X. Serra e J. O. Smith. “Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition”. In: *Computer Music Journal* 14.4 (1990), pp. 12–24.

[VRR06] F. Villavicencio, A. Röbel e X. Rodet. “Improving LPC spectral envelope extraction of voiced speech by true-envelope estimation”. In: *Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP 2006*. Vol. 1. IEEE. 2006.

[YR06] C. Yeh e A. Röbel. “Adaptive noise level estimation”. In: *Proc. of the 9th Int. Conf. on Digital Audio Effects, DAFx06*. 2006.